

# Semi-classical approach to sequential recombination algorithms for jet clustering

Jeff Tseng\* and Hannah Evans

*University of Oxford, Subdepartment of Particle Physics,  
Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, United Kingdom*

(Dated: April 9, 2013)

We derive a new sequential recombination algorithm for reconstructing jets of particles in high-energy collision events from a simple, semi-classical model of successive uniform massless emissions. The model results in a different distance measure used to determine the sequence of clustering steps, and effectively subtracts background as it reconstructs the jet. We examine the new algorithm's behavior in light of existing algorithms, and we find that in Monte Carlo comparisons, the new algorithm's robustness against collision backgrounds is comparable to that of other jet algorithms when the latter have been augmented by further background subtraction techniques.

Collimated jets of particles are a distinctive feature of high energy elementary particle collisions and are often taken to indicate the presence of ejected quarks or gluons, particles normally shrouded by the effects of quantum chromodynamics (QCD). Jet reconstruction therefore plays a prominent role in event analysis, and as the search for new physics breaches new thresholds in energy and jet multiplicity, understanding jet reconstruction itself has taken on new importance. This importance is especially true in the study of highly relativistic (“boosted”) objects, in which evidence of heavy or exotic particle production and decay can be discerned in a jet's substructure. Experimental results on jet substructure have been published by the CDF [1], ATLAS [2], and CMS [3] experiments.

A standard class of methods for jet reconstruction in hadron collider experiments is the sequential recombination algorithm. Different varieties of this algorithm usually are rooted in physical or geometric considerations, such as QCD splitting functions for the  $k_T$  algorithm [4, 5], angular ordering for Cambridge-Aachen [6], and collimated jet cores for anti- $k_T$  [7]. It is also possible to look at jets from the perspective of the relativistic boosts themselves. This perspective has been used to motivate, for instance, so-called “variable- $R$ ” jet algorithms [8], which focus on resonance decays within the jet. Successive gluon emissions in the jet, however, broaden the jet even further than would be expected from resonance decays alone.

In this article, we consider a simplified, semi-classical model based on relativistic boosts of these successive emissions. The model is used to derive a new sequential recombination algorithm which simultaneously removes background radiation, including initial state radiation as well as that originating from the unassociated collisions (“pileup”) which are an important feature of modern high-luminosity colliders such as the Large Hadron Collider. We test the new algorithm on simulated high-energy  $W$  bosons with and without the presence of pileup, and compare the results with those of other clustering algorithms. We will find that the new algorithm's performance is similar to that of the other al-

gorithms after further background subtraction (“grooming”) techniques have been applied to the latter. The new algorithm can be a useful addition to the experimenter's toolkit for resolving the structure of highly energetic collision products.

In the semi-classical model, we conceive of a parton-initiated jet as a parent particle with some effective mass and defined energy and direction in the laboratory frame. This parent, and its children, undergo a series of relatively soft, massless emissions which are uniform in their own rest frames. Each emission is then boosted into the laboratory frame with the direct ancestor's remaining energy. The angular probability density in the laboratory frame is

$$f(\theta, \phi) = \frac{1}{4\pi\gamma^2(1 - \beta \cos \theta)^2} \quad (1)$$

where  $\theta$  is the angle between the emission and the direct ancestor's direction, and  $\gamma$  and  $\beta$  are the boost factors into the laboratory frame. The probability of finding an emission in solid angle  $d\Omega$  is then

$$f(\theta, \phi)d\Omega = \frac{\sin \theta d\theta d\phi}{4\pi\gamma^2(1 - \beta \cos \theta)^2}. \quad (2)$$

Non-classical effects, such as those from spin and color, are expected to scatter a small amount of non-spherical radiation [9, 10], and are neglected in this model. Gluon jets are also broader than those of quarks; as a result, in common with other generic jet algorithms, different initial partons may be reconstructed with different efficiencies.

Jet clustering can be thought of as choosing the most likely sequence of  $1 \rightarrow 2$  splittings to produce the observed jet. Maximizing the above probability at each step is the same as finding the smallest distance  $w_{ij}$  between pairs of pre-existing clusters,

$$w_{ij} = (E_i + E_j)^2 \frac{(1 - \cos \theta_{ij})^2}{\sin \theta_{ij}}, \quad (3)$$

where  $i$  and  $j$  are the clusters,  $\theta_{ij}$  the angle between them, and  $E_i + E_j$  the energy sum which takes the place

of  $\gamma$ . We have let  $\beta \approx 1$  for simplicity. We make the usual replacements, for the hadron collider environment, of energy  $E$  with transverse energy  $E_T$ , and  $\theta_{ij}$  with  $\Delta R_{ij} = \sqrt{(\Delta y_{ij})^2 + (\Delta \phi_{ij})^2}$ , where  $\Delta y_{ij}$  is the rapidity difference and  $\Delta \phi_{ij}$  is the difference in azimuthal angle. Expanding the sines and cosines then gives us the new distance measure

$$d_{ij} = \frac{1}{4}(E_{Ti} + E_{Tj})^2 \left( \frac{\Delta R_{ij}}{R} \right)^3 \quad (4)$$

where we have introduced the jet scale parameter  $R$ , analogous to that of other inclusive jet algorithms. The effect of the coefficient  $1/4$  is that when we define the cluster-beam distance measure

$$d_{iB} = E_{Ti}^2, \quad (5)$$

the result is that  $d_{iB} < d_{ij}$  whenever two jets with the same  $E_T$  are separated by  $\Delta R_{ij} > R$ . If  $E_{Ti} < E_{Tj}$ , we also have  $d_{iB} < d_{ij}$  whenever  $\Delta R_{ij} > R$ , and therefore  $R$  is, as in other algorithms, the maximum  $\Delta R_{ij}$  between clusters that can be merged.

We follow the usual steps for a sequential recombination algorithm: the distances  $d_{ij}$  are calculated for each pair of clusters  $i$  and  $j$ , and  $d_{iB}$  for each cluster  $i$ . If the smallest distance is a  $d_{ij}$ , the pair is merged by adding their 4-momenta. If the smallest distance is a  $d_{iB}$ , the cluster is deemed an independent jet and removed from further consideration. These steps are repeated until all clusters have been deemed jets. This “semi-classical” (SC) algorithm is collinear and infrared-safe by construction.

It is useful to compare the new algorithm with the inclusive  $k_T$  algorithm, which uses the pair distance measure

$$d_{ij} = \min[E_{Ti}^2, E_{Tj}^2] \left( \frac{\Delta R_{ij}}{R} \right)^2. \quad (6)$$

One obvious difference is the increased  $\Delta R_{ij}$  exponent. The  $k_T$  algorithm, in common with most other algorithms in general use, incorporate the factor  $\Delta R_{ij}^2$  or its close relative  $(1 - \cos \theta_{ij})$ . Overall, the effect of the different exponent is not dramatic: larger exponents have been tested, and, for the most part, merely allow clusters to merge with larger  $\Delta R_{ij}$ , increasing the reconstructed jet size. For the rest of this article, we keep the exponent as 3, as that directly motivated by the semi-classical model.

The different energy factor, on the other hand, changes the order in which clusters are merged and set aside as jets. The  $k_T$  algorithm starts by merging soft clusters, as one would expect for an algorithm which attempts to reverse the splitting history, but avoids the perceived problem of the JADE algorithm [11, 12], with distance measure  $d_{ij} = E_i E_j (1 - \cos \theta_{ij})$ , which can allow large angle clusterings of very soft pairs. The semi-classical algorithm also starts by merging soft pairs, though the raised

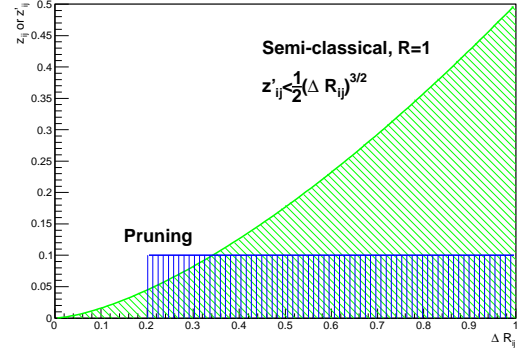


FIG. 1. Comparison of the semi-classical algorithm with pruning. The diagonally hashed region indicates mergings rejected by the semi-classical algorithm, while the vertically hashed region is for pruning. The pruning parameters are taken from [15].

$\Delta R_{ij}$  exponent clusters some high- $E_T$  clusters sooner if they are sufficiently close. Large angle clusterings are suppressed by the  $R$  scale and beam clustering. However, the most significant difference in behavior between the semi-classical and  $k_T$  algorithms is that for sufficiently large  $\Delta R_{ij}$  (though still with  $\Delta R_{ij} < R$ ), the comparison with  $d_{iB}$  prevents a number of soft clusters from merging with high- $E_T$  clusters when

$$z'_{ij} \equiv \frac{E_{Ti}}{E_{Ti} + E_{Tj}} < \frac{1}{2} \left( \frac{\Delta R_{ij}}{R} \right)^{3/2}. \quad (7)$$

As a result, while  $R$  defines the maximum extent of a jet in the semi-classical algorithm, the actual jets are likely to be narrower, with higher  $E_T$  associated with narrower jets. This behavior is similar to that of jet “pruning”, which vetoes mergings which satisfy the two conditions

$$z_{ij} \equiv \frac{\min(p_{Ti}, p_{Tj})}{|\vec{p}_{Ti} + \vec{p}_{Tj}|} < z_{cut}, \quad (8)$$

$$\Delta R_{ij} > D_{cut}, \quad (9)$$

and discards the softer of the two clusters [13, 14]. Figure 1 compares the two methods, with pruning removing the rectangular region in the  $(\Delta R_{ij}, z_{ij})$  plane, while the semi-classical algorithm additionally removes some soft clusters at small angles as well as harder clusters at large  $\Delta R_{ij}$ . As these clusters become stand-alone jets, it is possible for final jets to be separated by  $\Delta R_{ij} < R$ . This behavior follows from the algorithm’s underlying model of a single light parton jet, in which a large-angle split into two high-energy subjects is unlikely. Instead, most of the energy is assumed to be highly collimated in a single, narrow cluster.

Initial studies of the semi-classical algorithm with boosted objects have been performed using the PYTHIA (version 8.150) Monte Carlo generator [16, 17]. Single

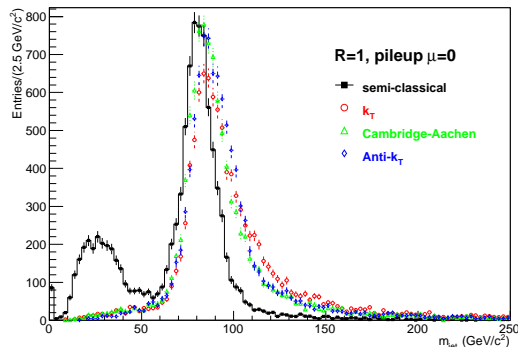


FIG. 2. Jet mass distributions for high- $p_T$  jets in the same hemisphere as the generated  $W$  boson, with no pileup and  $R = 1$  for the semi-classical,  $k_T$ , Cambridge-Aachen, and anti- $k_T$  algorithms.

hadronically decaying  $W$ +parton events were generated with  $W$   $p_T > 500$  GeV/ $c$  at  $\sqrt{s} = 8$  TeV. Non-neutrino particles were then collected into  $0.1 \times 0.1$   $\eta - \phi$  cells out to  $|\eta| < 5$ , where  $\eta = -\ln[\tan(\theta/2)]$  is pseudorapidity. Up to an average of 25 QCD minimum bias events, using Tune 4Cx [18] and the CTEQ6L1 parton distribution functions [19], were overlaid as “pileup”, assuming the same interaction vertex. Only cells with energy greater than 0.5 GeV were considered for jet clustering. Jets were then found using the  $k_T$ , Cambridge-Aachen, and anti- $k_T$  algorithms implemented in FASTJET version 3.0.3 [20], and the semi-classical algorithm implemented as a FASTJET plugin [21]. Jet masses were calculated by summing the 4-momenta of the cells, assuming zero mass for each cell.

Figure 2 shows the jet mass distribution for jets with  $p_T > 400$  GeV/ $c$  in the same hemisphere as the generated  $W$  for the different ungroomed jet algorithms with  $R = 1$ . Even with no pileup, the effect of additional radiation can be seen in the other algorithms, while the semi-classical peak is narrowest and lies closest, at  $80.9 \pm 0.1$  GeV/ $c^2$ , to the generated  $W$  mass of 80.385 GeV/ $c^2$ . The low and zero-mass bumps are the result of the semi-classical algorithm “pruning” close but energetically unbalanced  $W$  daughters, as noted above; combining the jet with another nearby jet recovers the  $W$  mass. When the pileup level increases to an average of 25, as shown in Figure 3, the semi-classical peak shifts roughly 4 GeV/ $c^2$  higher, but remains a recognizable, narrow peak, while the others are much broader due to incorporating pileup radiation.

The effects of additional radiation usually are mitigated by reducing the  $R$  parameter, and indeed one can see in Figure 4 that at  $R = 0.4$ , all the peak masses cluster around 80 GeV/ $c^2$ , rising rapidly for the other ungroomed algorithms. The semi-classical algorithm, on the other hand, starts low at  $R = 0.4$ , where the two  $W$  daughters often are resolved into different jets, and levels

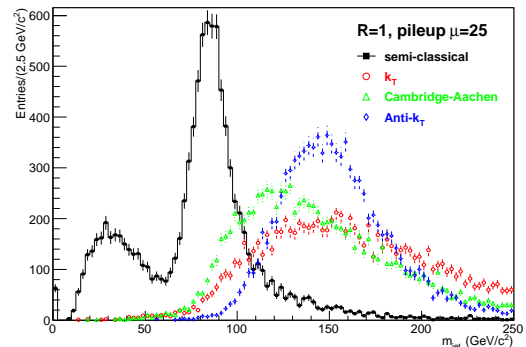


FIG. 3. Jet mass distributions for high- $p_T$  jets in the same hemisphere as the generated  $W$  boson, with an average of 25 pileup events overlaid and  $R = 1$  for the semi-classical,  $k_T$ , Cambridge-Aachen, and anti- $k_T$  algorithms.

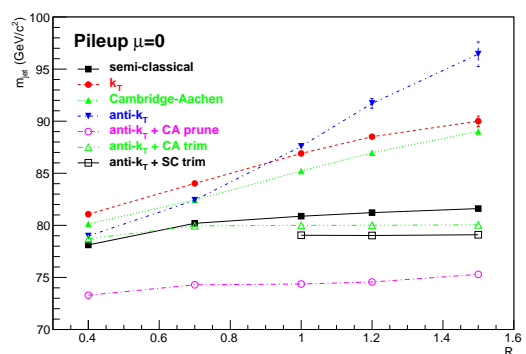


FIG. 4. Peak mass vs  $R$  for events with zero pileup.

off above  $R = 0.7$ .

Boosted object analyses, however, typically use large  $R$  values between 1 and 1.5 in order to remain sensitive to a larger range of energies. In order to mitigate pileup effects in such large jets, the jet can undergo further “grooming”. It is therefore instructive to compare the new algorithm with grooming techniques, several of which, including pruning, are also shown in Figures 4 and 5. It should be noted that grooming techniques usually are tailored to particular environments, and rely on knowledge of the target final state such as one might use to design a search strategy based on individually resolved jets. The comparisons shown in this article are therefore indicative, leaving optimization for specific signals and backgrounds for those particular analyses.

Pruning has already been described. We start with anti- $k_T$  jets with a given  $R$ , and use the parameters  $z_{cut} = 0.1$  and  $D_{cut} = 0.2$  [15] to prune. We compare the resulting jets with those from the semi-classical algorithm by itself, with the same  $R$ . Not surprisingly, the two algorithms behave similarly in Figures 4 and 5, even rising at a similar rate when the average pileup level is 25.

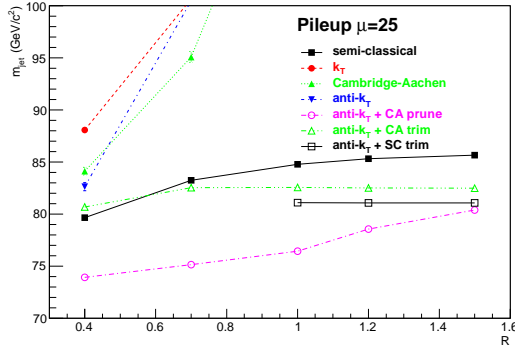


FIG. 5. Peak mass vs  $R$  for events with average 25 pileup. For most values of  $R$  for the  $k_T$ , Cambridge-Aachen, and anti- $k_T$  algorithms, the distributions are broad rather than peaked.

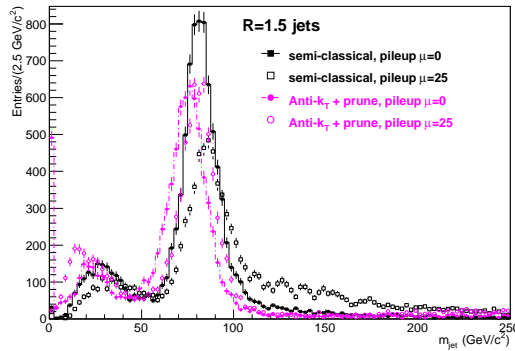


FIG. 6. Jet mass distributions for high- $p_T$  semi-classical and pruned anti- $k_T$  jets in the same hemisphere as the generated  $W$  boson, with zero and average 25 pileup.

Jet mass distributions for  $R = 1.5$  are shown in Figure 6. The presence of pileup shifts the peaks of the distributions upward, as expected. The semi-classical algorithm, however, leaves a larger high-mass tail, but also a smaller low-mass bump, suggesting that while it eliminates less pileup radiation, it retains both  $W$  daughters more often. It is also evident that the given pruning parameters are too aggressive for these particular conditions, resulting in a low peak mass.

Next, we consider the grooming technique of trimming, which attempts to discern narrow, high- $p_T$  subjets within the parent jet [22, 23]. For the comparison, we use the Cambridge-Aachen algorithm to recluster within the parent jet with a smaller radius parameter  $R_{sub} = 0.3$ , and discard the resulting subjets with  $p_T < f_{sub}P_T$ , where  $f_{sub} = 0.05$  is a parameter and  $P_T$  is the transverse momentum of the parent jet [15]. The jet mass is then calculated by summing the remaining high- $p_T$  subjets. Figure 5 shows trimming to be more stable under these pileup conditions than pruning or the ungroomed semi-classical algorithm.

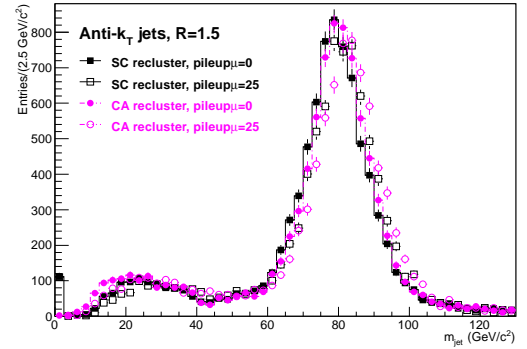


FIG. 7. Trimmed jet mass distributions for high- $p_T$  jets in the same hemisphere as the generated  $W$  boson, with zero and 25 pileup. Reclustering of the parent anti- $k_T$  jet has been performed using the Cambridge-Aachen algorithm with  $R_{sub} = 0.3$ , and the semi-classical algorithm with  $R_{sub} = 0.4$ .

The semi-classical algorithm can be used for reclustering, and indeed, reclustering is arguably a more natural context for the new algorithm than event-level clustering, as its underlying model is that of a single-parton jet, rather than a complex object incorporating two or more energetic decay products. In effect, this method combines pruning and conventional trimming. Figure 7 shows the results of reclustering with the Cambridge-Aachen and semi-classical algorithms. With the latter, we use a slightly larger value of  $R_{sub} = 0.4$  to compensate for the smaller semi-classical jets. Again, low-mass bumps are observed, where the other  $W$  daughter has been discarded by the trimming technique. As expected, the mass distributions are very similar, and are also largely insensitive to both the parent jet's  $R$  parameter and the pileup level.

Figures 8 and 9 show the effect of increasing the pileup level on the different ungroomed and groomed algorithms for two large  $R$  values. The difference between ungroomed and groomed jets is more obvious here, with the mass peak rapidly rising and broadening at even modest levels of pileup for all the ungroomed algorithms except the semi-classical algorithm. The ungroomed semi-classical algorithm parallels pruning over this range of pileup level, while trimming, with either reclustering algorithm, is more stable than pruning for this boosted  $W$  final state.

In this article, we have used a much simplified, semi-classical approach to motivate a new distance measure in a sequential recombination algorithm for jet clustering. The resulting algorithm effectively combines jet clustering with pruning-like behavior in one step. Monte Carlo tests with PYTHIA8 show the algorithm by itself performing like an algorithm with jet grooming in terms of stability with respect to the jet scale parameter  $R$  as well as to pileup. It can also be used to recluster narrow subjets

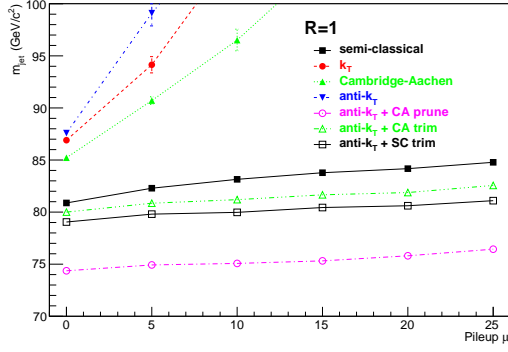


FIG. 8. Dependence of mass peak position on pileup for different algorithms with  $R = 1$ . The mass distributions at most pileup levels for the  $k_T$ , Cambridge-Aachen, and anti- $k_T$  algorithms are very broad, with maxima above  $100 \text{ GeV}/c^2$ .

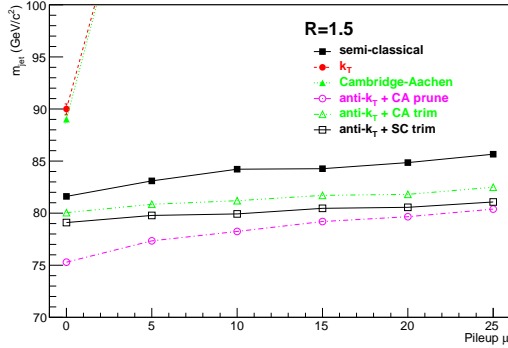


FIG. 9. Dependence of mass peak position on pileup for different algorithms with  $R = 1.5$ . The mass distributions at most or all pileup levels for the  $k_T$ , Cambridge-Aachen, and anti- $k_T$  algorithms are very broad, with maxima above  $100 \text{ GeV}/c^2$ . The anti- $k_T$  mass distribution peaks near  $100 \text{ GeV}/c^2$  even at zero pileup.

for trimming. Further work would be needed to determine whether cross sections can be calculated for the new algorithm without large QCD corrections. At the same time, as has been observed widely (and wisely), Monte Carlo studies may show the feasibility of a method, but they are a far cry from optimizing and testing it in a genuine experimental context.

This work was supported by the Science and Technology Facilities Council of the United Kingdom and the

Higher Education Funding Council of England. The authors would like to thank A. Cooper-Sarkar, C. Issever, and B.T. Huffman for useful comments and discussion.

\* j.tseng1@physics.ox.ac.uk

- [1] T. Aaltonen *et al.* (CDF Collaboration), Phys.Rev. **D85**, 091101 (2012), arXiv:1106.5952 [hep-ex]
- [2] G. Aad *et al.* (ATLAS Collaboration), JHEP **1205**, 128 (2012), arXiv:1203.4606 [hep-ex]
- [3] S. Chatrchyan *et al.* (CMS Collaboration)(2013), arXiv:1303.4811 [hep-ex]
- [4] S. Catani, Y. L. Dokshitzer, M. Seymour, and B. Webber, Nucl.Phys. **B406**, 187 (1993)
- [5] S. D. Ellis and D. E. Soper, Phys.Rev. **D48**, 3160 (1993), arXiv:hep-ph/9305266 [hep-ph]
- [6] M. Wobisch and T. Wengler(1998), arXiv:hep-ph/9907280 [hep-ph]
- [7] M. Cacciari, G. P. Salam, and G. Soyez, JHEP **0804**, 063 (2008), arXiv:0802.1189 [hep-ph]
- [8] D. Krohn, J. Thaler, and L.-T. Wang, JHEP **0906**, 059 (2009), arXiv:0903.0392 [hep-ph]
- [9] V. M. Abazov *et al.* (D0 Collaboration), Phys.Rev. **D83**, 092002 (2011), arXiv:1101.0648 [hep-ex]
- [10] D. Curtin, R. Essig, and B. Shuve(2012), arXiv:1210.5523 [hep-ph]
- [11] W. Bartel *et al.* (JADE Collaboration), Z.Phys. **C33**, 23 (1986)
- [12] S. Bethke *et al.* (JADE Collaboration), Phys.Lett. **B213**, 235 (1988)
- [13] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, Phys.Rev. **D80**, 051501 (2009), arXiv:0903.5081 [hep-ph]
- [14] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, Phys.Rev. **D81**, 094023 (2010), arXiv:0912.0033 [hep-ph]
- [15] (2012)
- [16] T. Sjostrand, S. Mrenna, and P. Z. Skands, JHEP **0605**, 026 (2006), arXiv:hep-ph/0603175 [hep-ph]
- [17] T. Sjostrand, S. Mrenna, and P. Z. Skands, Comput.Phys.Commun. **178**, 852 (2008), arXiv:0710.3820 [hep-ph]
- [18] R. Corke and T. Sjostrand, JHEP **1105**, 009 (2011), arXiv:1101.5953 [hep-ph]
- [19] J. Pumplin, D. Stump, J. Huston, H. Lai, P. M. Nadolsky, *et al.*, JHEP **0207**, 012 (2002), arXiv:hep-ph/0201195 [hep-ph]
- [20] M. Cacciari and G. P. Salam, Phys.Lett. **B641**, 57 (2006), arXiv:hep-ph/0512210 [hep-ph]
- [21] <http://www-pnp.physics.ox.ac.uk/~tseng/scplugin/>
- [22] D. Krohn, J. Thaler, and L.-T. Wang, JHEP **1002**, 084 (2010), arXiv:0912.1342 [hep-ph]
- [23] A. Abdesselam, E. B. Kuutmann, U. Bitenc, G. Brooijmans, J. Butterworth, *et al.*, Eur.Phys.J. **C71**, 1661 (2011), arXiv:1012.5412 [hep-ph]